Population transcriptomics reveals weak parallel genetic basis in repeated marine and freshwater divergence in nine-spined sticklebacks

Wang, Yingnan

2020-05

Wang , Y , Zhao , Y , Wang , Y , Li , Z , Guo , B & Merilä , J 2020 , ' Population transcriptomics reveals weak parallel genetic basis in repeated marine and freshwater divergence in nine-spined sticklebacks ' , Molecular Ecology , vol. 29 , no. 9 , pp. 1642-1656 . https://doi.org/10.1111/mec.15435

http://hdl.handle.net/10138/328958
https://doi.org/10.1111/mec.15435

unspecified
acceptedVersion

DR. ZITONG LI (Orcid ID : 0000-0001-8469-7295)

DR. BAOCHENG GUO (Orcid ID : 0000-0002-9796-5700)

PROF. JUHA MERILÄ (Orcid ID : 0000-0001-9614-0072)

Article type : Original Article

Population transcriptomics reveals weak parallel genetic basis in repeated marine and freshwater divergence in nine-spined sticklebacks

Yingnan Wang[1,3], Yongxin Zhao[1†], Yu Wang[1,3], Zitong Li[2], Baocheng Guo[1,3,4,*], Juha Merilä[2]

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

[2]Ecological Genetics Research Unit, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, FI-00014 University of Helsinki, Finland

[3]University of Chinese Academy of Sciences, Beijing 100049, China

[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

[†]Current address: Beijing Agro-biotechnology Research Center, Beijing Academy of Agriculture and Forestry Science, Beijing 100097, China

[*]**Corresponding author:** Baocheng Guo, Tel: +86 1064807978, E-mail: guobaocheng@ioz.ac.cn

Keywords: Gasterosteidae, Genetic parallelism, RNA-seq, SNP, Expression divergence

Running title: Population transcriptomics in nine-spined sticklebacks

Abstract

The degree to which adaptation to similar selection pressures is underlain by parallel *vs.* non-parallel genetic changes is a topic of broad interest in contemporary evolutionary biology. Sticklebacks provide opportunities to characterize and compare the genetic underpinnings of repeated marine-freshwater divergences at both intra- and interspecific levels. While the degree of genetic parallelism in repeated marine-freshwater divergences has been frequently studied in the three-spined stickleback (*Gasterosteus aculeatus*), much less is known about this in other stickleback species. Using a population transcriptomic approach, we identified both genetic and gene expression variations associated with marine-freshwater divergence in the nine-spined stickleback (*Pungitius pungitius*). Specifically, we used a genome-wide association study approach, and found that ~1% of the total 173,491 identified SNPs showed marine-freshwater ecotypic differentiation. A total of 861 genes were identified to have SNPs associated with marine-freshwater divergence in nine-spined stickleback, but only 12 of these genes have also been reported as candidates associated with marine-freshwater divergence in the three-spined stickleback. Hence, our results indicate a low degree of interspecific genetic parallelism in marine-freshwater divergence. Moreover, 1,578 genes in the brain and 1,050 genes in the liver were differentially expressed between marine and freshwater nine-spined sticklebacks, ~5% of which have also been identified as candidates associated with marine-freshwater divergence in the three-spined stickleback. However, only few of these (e.g., *CLDND1*) appear to have been involved in repeated marine-freshwater divergence in nine-spined sticklebacks. Taken together, the results indicate a low degree of genetic parallelism in repeated marine-freshwater divergence both at intra- and interspecific levels.

Introduction

The topic of convergent evolution – the independent evolution of similar phenotypes at intraspecific and/or interspecific levels – has a long history in evolutionary biology (Darwin, 1859). Uncovering the molecular basis of convergence allows us to understand if organisms adopt the same/similar or different genetic solutions towards reaching the same phenotype, thus allowing us to address the degree to which evolution at the genetic level is repeatable and predictable (Rosenblum, Parent, & Brandt, 2014). There are many examples of genetic parallelism underlying

convergent evolution. For example, despite millions of years of divergence between bamboo-eating giant and red pandas, limb development genes *DYNC2H1* and *PCNT* appear to be important candidates for pseudothumb development in both species (Hu et al., 2017). Similarly, evolutionarily independent fish lineages leverage similar transcription factors, developmental and cellular pathways in evolving electric organs (Gallant et al., 2014). Likewise, the fatty acid desaturase gene *Fads2* is suggested to play a key role in facilitating recurrent freshwater colonization in fishes (Ishikawa et al., 2019).

In the same vein, the repeated loss of the pelvic apparatus in three-spined sticklebacks (*Gasterosteus aculeatus*) provides an example of genetic parallelism based on recurrent deletion of a *Pitx1* enhancer due to its sequence fragility (Chan et al., 2010; Peichel et al., 2001; Shapiro et al., 2004; Shapiro, Marks, et al., 2006; Xie et al., 2019). In fact, a number of genomic regions have been identified to be consistently associated with marine-freshwater divergence in three-spined sticklebacks (Hohenlohe et al., 2010; Jones et al., 2012; Jones et al., 2012). While all of these studies have looked at genetic variation across the genome, changes associated with marine-freshwater divergence across the transcriptome have been less studied (but see Gibbons, Metzger, Healy, & Schulte, 2017; Ishikawa et al., 2017; Jones et al., 2012; Kusakabe et al., 2017; Pritchard et al., 2017; Wang et al., 2014).

Changes in gene expression have long been suspected to underlie biological functions and phenotypic diversity (King & Wilson, 1975), and to also play key roles in convergent evolution (Ogura, Ikeo, & Gojobori, 2004). For example, convergent origins of complex bioluminescent organs in squids have been found to be associated with widespread parallel changes in gene expression (Pankey, Minin, Imholte, Suchard, & Oakley, 2014). Similarly, parallel expression shifts are observed in response to high-altitude environmental stresses in birds (Hao et al., 2019). The major argument for the role of gene expression differentiation came from the realization that many genes have tissue-specific enhancer elements, and changes in these would be expected to have fewer pleiotropic effects on gene function than changes in protein coding sequences (Carroll, 2005).

Changes in gene expression seem particularly relevant in the context of repeated marine-freshwater divergence in three-spined sticklebacks. The changes in regulatory sequences appear to predominate those in coding sequences in the set of genomic regions associated with repeated marine-freshwater divergence (Jones et al., 2012). Accordingly, the genome-wide landscape of

gene expression divergence between marine and freshwater three-spined sticklebacks has been further investigated, and many candidate genes whose expression is associated with salinity tolerance have been identified (Gibbons et al., 2017; Ishikawa et al., 2017; Kusakabe et al., 2017; Wang et al., 2014). However, earlier genome-wide studies of gene expression in three-spined sticklebacks have usually been based on pairwise comparisons of a single marine-freshwater population pair, and thus, the results may be confounded by population-specific divergence. In addition, although expression of certain genes has been independently suggested to be associated with marine-freshwater adaptation in three-spined sticklebacks, it is not known if parallel divergence in expression is confined to a similar set of genes in repeated marine-freshwater divergences in other stickleback species that also inhabit marine and freshwater habitats, such as the *Pungitius pungitius* (Wootton, 1976).

Similar to three-spined sticklebacks, nine-spined sticklebacks have repeatedly and independently evolved similar morphological (e.g., pelvic apparatus and lateral plate reduction), behavioral, neuroanatomical, and physiological phenotypes in response to life in freshwater (Merilä, 2013). Given that the two stickleback species diverged about 26 million years ago (Fang et al., 2019; Varadharajan et al., 2019), they offer an excellent opportunity to study repeated marine-freshwater divergence at both intra- and interspecific levels. Compared to three-spined sticklebacks, population genomic studies of nine-spined sticklebacks are still rare (Bruneaux et al., 2013; Guo, Chain, Bornberg-Bauer, Leder, & Merilä, 2013; Raeymaekers et al., 2017; Varadharjan et al., 2019; Guo et al., 2019; Li, Löytynoja, Fraimout, & Merilä, 2019). To this end, we took a population transcriptomic approach to study both genetic and gene expression divergence between marine and freshwater populations of nine-spined sticklebacks. Specifically, we aimed to determine how frequently the same genetic and gene expression changes are associated with repeated marine-freshwater divergences in nine-spined sticklebacks, and if these changes are similar to those associated with marine-freshwater divergence in three-spined sticklebacks.

Materials and Methods

Sample collection

Adult nine-spined sticklebacks were collected during the breeding season (June-July) of 2013 from six Fennoscandian sites, including two marine and four freshwater populations (Fig. 1; Table S1). The fish were captured with hand seines and/or minnow traps (mesh size 6 mm), and transported to the laboratory in Helsinki where they were allowed to stabilize in freshwater at

17°C under a 14h light:10h dark photoperiod for seven days. During this time, the fish were fed twice per day with chopped chironomid larvae. From each of the six populations, two males and two females (n = 24) were randomly selected for dissection. Tissues, brains and livers were dissected and immediately frozen in liquid nitrogen; they were later transferred to -80°C, where they were maintained until RNA-extraction.

RNA extraction and sequencing

In order to access a large number of transcripts, the transcriptomes of two highly complex organs – brain and liver – from each of the 24 individuals were sequenced. Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's protocol. cDNA libraries and sequencing were done by BGIHONGKONG CO., LIMITED. Briefly, magnetic beads with Oligo (dT) were used for isolating mRNA after DNase I treatment on total RNA; the mRNA was fragmented into short fragments and cDNA was synthesized using these fragments as templates. The short cDNA fragments were purified and ligated to adapters; 200 bp fragments were selected for PCR amplification, and the products were sequenced on the Illumina HiSeq2000 platform with 90 bp paired-end strategy. Each of the 48 samples were sequenced twice on two different sequencing lanes to obtain technical replicates. In total, 1.26 billion reads of the 48 transcriptomes were obtained. The number of reads for each transcriptome ranged from 10.5 to 34.4 million, and from 40.5 to 59.5 million for each individual (Table S1).

Read mapping

The nine-spined stickleback genome (Varadharajan et al., 2019) was used as the reference genome, which includes 21 pseudochromosomes (hereinafter referred to as chromosomes), as well as the three-spined stickleback genome used in Rastas, Calboli, Guo, Shikano, & Merilä (2016). Quality filtered reads from each sequenced transcriptome were aligned to the reference genome using HISAT2 version 2.0.1 (Pertea, Kim, Pertea, Leek, & Salzberg, 2016) with default settings, incorporating known gene annotations. The mapping results were converted from SAM to BAM format using SAMtools version 1.4 (Li et al., 2009). Although duplicates identified during alignment might sometimes be true biological signals, the probability of bias due to removal of wrong reads is greatly reduced when using paired-end sequencing (Parekh, Ziegenhain, Vieth, Enard, & Hellmann, 2016). Thus, sorted and duplicate-removed BAM format mapping results were used in the analyses of genetic and expression differentiation. In total, 72.5% of the reads (0.9 billion) were aligned to the reference genome. The percentage of reads that were aligned to

the reference genome ranged between 60.6 and 86.0 for each transcriptome, and between 63.8 and 78.0 for each individual.

SNP detection and annotation

Single Nucleotide Polymorphisms (SNPs) were identified as follows: BAM format mapping results from the same individual were first merged, and SNPs were then called across the six populations with mapping quality ≥ 20 using 'mpileup' in SAMtools and BCFtools. SNPs with base coverage of DP < 100 or DP > 1,000 per individual, missing genotype in more than four individuals, and minor allele frequency < 0.05 across all samples were excluded. Finally, only biallelic SNPs were kept for the analyses. SNPs were annotated (e.g., coding *vs.* non-coding, and synonymous *vs.* nonsynonymous) using the latest version of ANNOVAR (2019Oct24) (Yang & Wang, 2015).

Analyses of population structure

Autosomal SNPs were used for investigating genetic relationships among the six study populations by excluding SNPs in chromosome 12 – the nine-spined stickleback sex chromosome (Rastas et al., 2016; Shapiro et al., 2009) – and in the unassembled scaffolds (Table S2). Matrices for principal component analysis (PCA) and neighbour-joining (NJ) tree estimation were obtained using PLINK version 1.9 (Purcell et al., 2007). Model-based clustering was performed using STRUCTURE version 2.3 (Pritchard, Stephens, & Donnelly, 2000), considering values of K from 2 to 6 with 10 independent runs for each K value, 15,000 burnin, and 35,000 simulation cycles. The average cluster membership of each K value was calculated using CLUMPP version 1.1.2 (Jakobsson & Rosenberg, 2007), and then visualized via DISTRUCT version 1.1 (Rosenberg, 2004). The most suitable value of population number (*K*) was inferred with STRUCTURE HARVESTER Web version 0.6.94 (Earl & Vonholdt, 2012).

Identification of SNPs associated with marine-freshwater differentiation

A single-locus genome-wide association approach (GWAS) was applied to identify SNPs that were associated with the marine-freshwater divergence in nine-spined sticklebacks. Habitat (marine *vs.* freshwater) was considered as a 'binary trait', and the logistic regression (Balding, 2006) was applied to identify SNPs that were significantly associated with habitat. The logistic regression is specified as

$$\min_{\beta} \sum_{i=1}^{n} [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad (1)$$

where

$$p_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j)}, \quad (2)$$

,

y_i (y_i = 0 for marine; y_i = 1 for freshwater) is the phenotype of individual i (i = 1 to n; n is the total number of individuals), x_{ij} (x_{ij} = 0 for genotype AA, x_{ij} = 1 for genotype AB, and x_{ij} = 2 for genotype BB) is the genotype of SNP j (j = 1 to p; p is the total number of SNPs) of individual i, β_0 is the parameter of the intercept or population mean, β_j is the effect of SNP j. The P-values of each SNP were evaluated. A permutation test was used for multiplicity adjustment to control false positives, which was conducted in a standard way as developed for regression-based association analysis (Foulkes 2009). The phenotype records were randomly reshuffled thousands of times, and for each replicate the association mapping was conducted to obtain a test statistic for each SNP. As such, an empirical distribution of test-statistics for each SNP was obtained, and then used for multiple hypothesis testing. To reduce the effects of missing data, the identified biallelic SNPs with a missing rate higher than 0.1 across all of the six studied populations were further excluded. The missing data were simply imputed by the mean value of the known genotypes at a given SNP. Since the sex effects were not significant ($P > 0.05$), sex was not included into the model as a factor.

Differential gene expression analysis

The gene expression profile in each individual, as well as the differentially expressed genes/transcripts (DEGs/DETs) in the population comparisons were characterized with the transcript-level expression analysis pipeline of HISAT, StingTie, and Ballgown for RNA-seq data (Pertea et al., 2016). In all DEGs/DETs analyses, both brain and liver data were used independently. Briefly, RNA-seq reads were first aligned to the nine-spined stickleback genome (Varadharajan et al., 2019) using HISAT2 version 2.0.1 to identify their genomic positions. Transcripts were then assembled and quantified for each of the 96 transcriptomes using StringTie version 1.2.2 with default parameters, merging with reference gene models. The expression of each transcript was quantified as Transcripts Per Million (TPM). TPM is known to be preferred

over Reads Per Kilobase Million (RPKM) and Fragments Per Kilobase Million (FPKM) for quantification, because it is independent of the mean expressed transcript length, and thus, more comparable across samples. After technical repeatability analysis, transcripts were assembled and quantified for each tissue of each individual using the merged BAM format mapping results from the technical replicates. A PCA was also conducted on the transcript data with TPM ≥ 1 using the default R function "prcomp()", which could reveal the population structure among the samples on the basis of gene expression. Considering that the DEGs/DETs identification might result from data heterogeneity, overall transcriptomic similarity between each pair of individuals was first evaluated both within and across populations by quantifying cosine similarity using the R package lsa, as described in Pankey et al. (2014). Finally, DEGs and DETs were identified in population comparisons using the R package Ballgown version 2.16.0 (Fu, Frazee, Collado-Torres, Jaffe, & Leek, 2019). Phenotypic data (*viz.* population, ecotype, and sex) for each individual was loaded to Ballgown, and sex was set as a covariate in each comparison. Transcripts with a variance of less than one across samples were excluded. A q value < 0.05 of false discovery rate (FDR) was used for identifying DEGs/DETs in population comparisons. It is worth noting that transcripts are more abundant than genes within a given tissue, which suggests that the criteria of DET identification is more stringent than that of DEGs when using the same q value for FDR.

Technical reproducibility of each of the 48 transcriptomes was high ($r_s \geq 0.96$, $P < 0.01$; Fig. S1), suggesting that the identified DEGs/DETs are not likely to be affected by data quality. DEGs/DETs potentially associated with marine-freshwater divergence in the studied populations were identified with a pooled approach (Berner & Salzburger, 2015), in which gene expression profiles are compared quantitatively between the two ecotype groups. In our case, the marine ecotype group included samples from the two marine populations, and the freshwater ecotype group were those from the four freshwater populations. Possible interactions among the identified DEGs/DETs associated with marine-freshwater divergence were constructed using a sparse (inverse) covariance matrix estimation approach proposed by Meinshausen & Bühlmann (2006). The idea is to take each gene in turn as the response variable, and all other genes as the explanatory variables to build a normal Elastic net regression model (Zou and Hastie 2005):

$$\frac{1}{2n}(x_{ik} - \sum_{j \neq k} x_{ij})^2 + \lambda[w\sum_{j \neq k}|\beta_k| + (1-w)\sum_{j \neq k}\beta_k^2]. \quad (3)$$

The Elastic net model can detect a subset of genes having a non-zero effect, meaning that those

genes are connected to the target gene. In the network, each gene was considered as a vertex, and if there is an association (i.e. non-zero regression coefficient) between a pair of vertexes, an edge is added between the two vertexes. The community structure of the constructed protein network was explored using the R package igraph. Specifically, the function membership() was applied to define clusters of vertexes based on their level of connectivity.

DEGs/DETs that might underlie repeated marine-freshwater divergence were further identified with an integrated approach, in which gene expression profiles between pairwise marine-freshwater populations were compared. To reduce false positives from habitat-unrelated differentiation, expression difference in comparisons of pairwise marine-marine and freshwater-freshwater populations were further considered (Berner & Salzburger, 2015). Here, a gene/transcript would not be identified as a DEG/DET associated with repeated marine-freshwater divergence in the studied populations if it was expressed differentially in either marine-marine or freshwater-freshwater comparisons, or only in one pairwise marine-freshwater comparison.

Gene Ontology enrichment

SNPs that were identified to be associated with marine-freshwater divergence were annotated using BEDTools 2.17.0 (Quinlan & Hall, 2010) to characterize the genes that they were located in. Identified DEGs/DETs were annotated using BEDTools 2.17.0 to obtain gene names, gene functions, and Gene Ontology (GO) terms from the annotated nine-spined stickleback genome (Varadharajan et al., 2019). GO enrichment analysis was conducted to test whether genes with SNPs associated with marine-freshwater divergence were significantly enriched for certain GO terms with the R package ClusterProfiler (Yu, Wang, Han, & He, 2012).

Results

Population genetic structure in nine-spined sticklebacks

In total, 233,343 biallelic SNPs were identified. The number of SNPs identified on each chromosome was significantly and positively correlated with chromosome length (r_s = 0.78, P = 3.36×10^{-5}; Fig. 2; Table S2). Of the 233,343 biallelic SNPs, 81,851 were in protein coding regions in 15,754 genes and 43,833 of these corresponded to synonymous changes and 38,018 to nonsynonymous changes, whereas 151,492 resided in non-coding regions. Of the 233,343 biallelic SNPs, 1,757 were specific to population FIN-HEL, 454 to FIN-PYO, 5,543 to FIN-RYT, 2,904 to SWE-ABB, 1,506 to SWE-BOL, and 2,454 to SWE-BYN. With the 233,343 biallelic SNPs, three

distinct population clusters were consistently observed in the PCA, NJ tree, and STRUCTURE analyses (Fig. 1). These corresponded to a cluster including the two marine populations (SWE-BOL and FIN-HEL), a cluster with the two Finnish freshwater populations (FIN-PYO and FIN-RYT), and a cluster including the two Swedish freshwater populations (SWE-BYN and SWE-ABB; Fig. 1). Within each of the clusters, divergence between freshwater populations is higher than that between marine populations, with each freshwater population forming a distinctive cluster and the two marine population clustering together (Fig. 1).

Candidate loci associated with marine-freshwater divergence in nine-spined sticklebacks

A subset of 173,491 bi-allelic SNPs with a missing rate not higher than 0.1 across all of the six populations were used in GWAS (Table S2). A total of 1,969 SNPs were identified to be associated with marine-freshwater divergence. The number of these SNPs on each chromosome was not correlated with chromosome length (r_s = 0.34, P = 0.13; Fig. 2; Table S2). Of the 1,969 SNPs, 1,429 were located in 861 genes on the 21 nine-spined stickleback chromosomes (Fig. 2). The number of genes harboring candidate SNPs on a given chromosome was significantly and positively correlated with chromosome length (r_s = 0.47, P = 0.03; Table S2). Of the 861 genes, 601 harbored only one candidate SNP, whereas 260 had ≥ 2 candidate SNPs (Table S3). Of the 861 genes, 223 harbored candidate SNPs with nonsynonymous changes, and 32 with ≥ 2 candidate SNPs with nonsynonymous changes. GO enrichment analysis showed that these 861 genes were significantly (P < 0.05) enriched in two GO terms: binding and ion binding (Fig. S2).

Differentially expressed genes/transcripts between marine and freshwater nine-spined sticklebacks

The gene expression PCA shows expression divergence between marine and freshwater nine-spined sticklebacks in the brain along PC2, which explains only 1% variation. No expression divergence was found between marine and freshwater nine-spined sticklebacks in the liver (Fig. S3). The overall transcriptomic similarity between each pair of individuals were 90% in the brain and 70% in the liver (Fig. S4).

With the pooled approach, 1,578 DEGs (out of 15,209 genes) and 1,373 DETs (out of 37,562 transcripts) were identified in the brain, and 1,050 DEGs (out of 13,599 genes) and 759 DETs (out of 29,335 transcripts) in the liver, between marine and freshwater nine-spined sticklebacks (Table S4). DEGs were found in all 21 chromosomes (Fig. 2). DEGs were enriched in 18 GO terms in the brain, and in 74 GO terms in the liver, whereas DETs were enriched in 22 GO terms in the brain,

and in 72 in the liver ($P < 0.05$; Fig. 3). 57 DEGs and 67 DETs were found in both brain and liver (Fig. S5). Interactions among identified DEGs associated with marine-freshwater divergence were found in the brain but not in the liver. In the brain, 11 clusters include ≥ 10 genes showing interactions among each other (Table S5). Six of the 11 clusters include DEGs identified with the pooled approach in the brain, and one cluster with 63 DEGs (Fig. S6).

With the integrated approach, 638 DEGs and 801 DETs were unique to pairwise marine-freshwater comparisons in the brain; 679 DEGs and 228 DETs were unique to pairwise marine-freshwater comparison in the liver (Fig. S7; Table S6). DEGs were enriched in 16 GO terms in the brain, and in 50 in the liver. DETs were enriched in 10 GO terms in the brain, and in two in the liver ($P < 0.05$; Fig. S8). Most of DEGs/DETs unique to pairwise marine-freshwater comparisons in the brain/liver appear in only one pairwise marine-freshwater comparison (Table 1). Only 14 DEGs (of 638) and 12 DETs (of 801) in the brain repeatedly appear in two or more pairwise marine-freshwater comparisons (Table 1). Similarly, only one DEG (of 679) and no DETs (of 228) were common to two or more comparisons in the liver data (Table 1).

Fifty-nine DEGs and 83 DETs identified with the pooled approach were found to be unique to a certain pairwise marine-freshwater comparison with the integrated approach in the brain. For the liver, the corresponding amounts were 32 DEGs and nine DETs. Thirteen DEGs and 12 DETs identified with the pooled approach are among those DEGs/DETs potentially associated with repeated marine-freshwater divergence in the brain, and one DEG (but no DET) in the liver. Fifteen DEGs identified with the pooled approach in the brain and three in the liver were found to harbor SNPs associated with marine-freshwater divergence in GWAS analyses reported above. Two DEGs were found in both the brain and the liver. These DEGs/DETs are hence highly likely to be associated with repeated marine-freshwater divergence in nine-spined sticklebacks; see Table 2 for their known annotations.

Discussion

The key finding of this study is the low degree of parallelism in gene expression differentiation associated with repeated marine-freshwater divergence in Northern European nine-spined stickleback populations. This suggests that the genetic underpinnings of adaptation to similar environments resulting from similar selection pressures could be very different, even in closely related populations. Nevertheless, several genes (1.74% of all divergent genes) were identified to be associated with repeated marine-freshwater divergence in the nine-spined stickleback with high

confidence. This confidence is based on the facts that divergent genes harbor functionally important amino acid substitutions, and that they are differentially expressed between marine and freshwater nine-spined stickleback populations. In the following, we will discuss these findings in light of repeatability of evolution, and genetic parallelism in freshwater adaptation in sticklebacks in particular.

Parallelism in genetic variation in repeated marine-freshwater divergence

Convergent evolution is often underlain by parallelism at the genetic level (Rosenblum et al., 2014; Stern, 2013), as often seen in the case of marine-freshwater divergence in three-spined sticklebacks (Jones et al., 2012). The underlying explanation for such genetic parallelism is that mutations in some particular genetic loci minimize pleiotropic effects while simultaneously maximizing the likelihood of adaptation (Stern, 2013). However, earlier studies indicate that the convergent evolution in nine-spined sticklebacks could be sometimes – even frequently (Merilä 2013, 2014) – based on non-parallel genetic changes. For example, similar to three-spined sticklebacks, marine nine-spined stickleback populations have fully developed pelvic apparatuses, whereas some freshwater populations display pelvic reduction (Blouw & Boyd, 1992; Herczeg, Turtiainen, & Merilä, 2010; Klepaker, Ostbye, & Bell, 2013). The *Pitx1* gene has been identified to be responsible for all known cases of pelvic reduction in the three-spined stickleback (Jones et al., 2012). However, in the case of nine-spined sticklebacks, it was identified as a major cause for the pelvic reduction in only one Canadian (Shapiro, Bell, & Kingsley, 2006) and one Finnish (Shikano, Laine, Herczeg, Vilkki, & Merilä, 2013) population, but not in several others (Shapiro et al., 2009; Kemppainen et al., 2020). In line with these findings, the results of the current study suggest that genetic changes associated with repeated marine-freshwater divergence in the nine-spined stickleback seem to be very different from those in the three-spined stickleback. Of the 861 genes with SNPs associated with marine-freshwater divergence in the nine-spined stickleback, only 12 were identified as candidate genes in marine-freshwater divergence in the three-spined stickleback (Table S3; Ferchaud et al., 2014; Hohenlohe et al., 2010; Jones, Chan, et al., 2012; Jones et al., 2012). This suggests that genetic changes associated with repeated marine-freshwater divergence in the two geographically coexisting and ecologically similar stickleback species are largely species specific and non-parallel.

After investigating genomic divergence between coexisting nine- and three-spined stickleback populations from the North Sea region, Raeymaekers et al. (2017) suggested that genomic

architecture, gene flow, and life history may collectively contribute to such differences between the two stickleback species. Rosenblum et al. (2014) highlighted that population size may strongly affect the probability of parallelism by influencing the dynamics of genetic drift, natural selection, and mutation. Because the role of chance in allele frequency change is more pronounced in small than in large populations, natural selection is less efficient in fixing beneficial mutations in small populations (Rosenblum et al., 2014; see also: Merilä 2013, 2014). In small populations, wherein founder events and random genetic drift prevail, potentially advantageous rare alleles (if even present within the founder groups) may be lost, and/or adaptation to given selection pressures might be more easily gained by allelic substitutions in alternate loci influencing the same polygenic trait (Merilä, 2013). Marine-freshwater divergence is likely to involve natural selection not only on genes coding for morphological traits, but also for genes involved in physiologically important functions, such as osmoregulation, thermal tolerance, and growth – many of which are known to have a polygenic basis (Healy, Brennan, Whitehead, & Schulte, 2018; Kusakabe et al., 2017; Laine, Shikano, Herczeg, Vilkki, & Merilä, 2013). The freshwater nine-spined stickleback populations studied here are known to be small, based on their very low genetic diversity (Merilä, 2013; Shikano, Shimada, Herczeg, & Merilä, 2010). As such, it is not surprising that nine-spined sticklebacks have adopted different genetic changes for repeated marine-freshwater divergence as compared to three-spined sticklebacks, whose Fennoscandian freshwater populations are typically much larger than those of the nine-spined sticklebacks (DeFaveri, Shikano, Ab Ghani, & Merilä, 2012). Thus, differences in effective population size, together with the polygenic nature of marine-freshwater divergence, could explain the differences in marine-freshwater divergence between the two stickleback species. However, it is also worth noting that the low degree of genetic parallelism associated with repeated marine-freshwater divergence between nine- and three-spined sticklebacks observed here is based on transcriptomic data, rather than whole genome resequencing data in nine-spined sticklebacks. A global marine and freshwater population comparison based on whole genome resequencing in nine-spined sticklebacks, similar to that in the three-spined stickleback (Jones et al., 2012), would be needed to evaluate the prevalence of genome-wide genetic parallelism – or lack thereof (see: Fang, Kemppainen, Momigliano, & Merilä, 2020) – between the two stickleback species.

Parallelism in gene expression in repeated marine-freshwater divergence

Gene expression variation might play a key role in the repeated marine-freshwater divergence in

three-spined sticklebacks: parallelism of regulatory changes predominates over coding changes (Jones, et al., 2012). For example, pelvic reduction is known to be underlain by variation in the cis-regulatory region of the *Pitx1* gene (Chan et al., 2010; Xie et al., 2019). In fact, a number of candidate genes whose expression variation is associated with marine-freshwater divergence in three-spined sticklebacks have been identified in diverse tissues using different methods (Gibbons et al., 2017; Ishikawa et al., 2017; Jones, et al., 2012; Kusakabe et al., 2017; Wang et al., 2014). Although the data are not always directly comparable among studies because different tissues have been used, some of the DEGs/DETs identified with the pooled approach between marine and freshwater nine-spined stickleback populations in this study are among the candidate genes of expression variation associated with marine-freshwater divergence in the three-spined stickleback (Table S4). However, these DEGs/DETs rarely showed repeated expression differentiation in pairwise comparisons between marine and freshwater nine-spined sticklebacks according to the integrated approach (Table 1). Interestingly, many DEGs/DETs in the pairwise comparisons between marine and freshwater nine-spined sticklebacks (according to the integrated approach) are also reported as candidate genes of expression variation associated with marine-freshwater divergence in the three-spined stickleback (Table S6). These results suggest that expression variation in some genes might be associated with marine-freshwater divergence in both nine- and three-spined sticklebacks, but parallel gene expression variation is rare in marine-freshwater divergence in the studied nine-spined stickleback populations. Identification of genes whose expression variation is associated with marine-freshwater divergence in sticklebacks requires multiple marine-freshwater comparisons to exclude population-specific effects, or usage of multiple approaches (Kusakabe et al., 2017). Leder et al. (2015) demonstrated substantial heritability of genome-wide gene expression variation in a three-spined stickleback population from the Baltic Sea. Likewise, the genetic basis of gene expression variation has been recently uncovered in several three-spined stickleback populations (Hart, Ellis, Eisen, & Miller, 2018; Pritchard et al., 2017). Notably, *trans* regulatory changes are predominant and more likely to be shared among convergently evolved populations, whereas different *cis* regulatory changes are more frequent in convergently evolved populations (Hart et al., 2018). Identification of genetic determinants of gene expression variation between marine and freshwater nine-spined stickleback populations would require expression quantitative trait loci mapping based on whole genome resequencing data. Such population genomic studies would also be useful to estimate the relative contribution of *trans* and *cis* regulatory changes underlying gene expression variation associated

with marine-freshwater divergence in the nine-spined stickleback.

It is also worth noting the difference in gene expression profiles between the two studied tissues. First, overall similarity in gene expression patterns in the brain was higher than that in the liver across all samples (Fig. S3 & S4). Second, gene expression differentiation associated with marine-freshwater divergence was more pronounced in the brain than in the liver. In addition to the overall higher differentiation in the brain (Fig. S3), the DEGs/DETs associated with repeated marine-freshwater divergence were also found mostly in the brain, and to a lesser extent in the liver, according to the multiple pairwise comparisons (Table 1). This observation is consistent with earlier studies that have found adaptive differentiation in brain size (Gonda, Herczeg, & Merilä, 2009; Gonda, Herczeg, & Merilä, 2011) and behavior (Herczeg, Gonda, & Merilä, 2009) between marine and freshwater nine-spined stickleback populations. Third, considering that expression profiles are typically tissue-specific (Brawand et al., 2011), it is not surprising that common DEGs/DETs between the brain and liver were rare in nine-spined sticklebacks when using the pooled method (Fig. S5). Taken together, our results suggest that transcriptomic comparisons of the brain, rather than liver, might better reflect gene expression differentiation associated with marine-freshwater divergence in nine-spined sticklebacks.

Finally, one methodological aspect relating to interpretation of gene expression results should be addressed. Given that we used standard RNA-seq libraries, the results might be subject to biases associated with removal of PCR duplicates: computational removal of PCR duplicates based only on their mapping coordinates are known to introduce biases into data analyses (Fu, Wu, Beane, Zamore, & Weng, 2018). However, paired-end sequencing (as used here) should reduce the likelihood of this bias. Such biases could be effectively eliminated by using unique molecular identifiers in RNA-seq library construction (Fu, Wu, Beane, Zamore, & Weng, 2018), as is now routinely done in single-cell RNA-seq studies (Stark, Grzelak, & Hadfield, 2019). This protocol can improve the accuracy of quantitative sequencing, and is now becoming more commonly used also in bulk RNA-seq studies (Stark, Grzelak, & Hadfield, 2019).

Candidate genes associated both genetic and expression parallelism in repeated marine-freshwater divergences

Although parallelism was rare in general, a number of genes were identified to be associated with repeated marine and freshwater divergence in nine-spined sticklebacks. Six genes were identified as DEGs/DETs with both the pooled and integrated approaches (Table 2), three of which have

been reported to be DEGs between marine and freshwater three-spined sticklebacks in different comparisons (Gibbons et al., 2017; Wang et al., 2014). Sixteen genes that were identified as DEGs/DETs with the pooled approach also harbored SNPs associated with marine-freshwater divergence in the GWAS analysis (Table 2), five of which have been identified to be associated with marine-freshwater divergence in earlier studies of three-spined sticklebacks. These genes are candidates for future functional validation. For example, the Claudin Domain Containing 1 (*CLDND1*) gene had seven SNPs associated with marine-freshwater divergence (Fig. 4A). One of these SNPs results in an amino acid change with Glutamine in marine nine-spined sticklebacks, and Lysine in freshwater nine-spined sticklebacks (Fig. 4B). Although the Glutamine-Lysine or Lysine-Glutamine change is predicted to be functionally tolerated (Vaser, Adusumalli, Leng, Sikic, & Ng, 2016), Glutamine and Lysine are different in many respects, e.g. potential side chain H-bonds, isoelectric point, hydrophobicity, etc. Interestingly, all SNPs occurred in the transmembrane domain of *CLDND1* protein. In addition, expression of *CLDND1* is significantly different between marine and freshwater nine-spined sticklebacks in both the brain and the liver (Fig. 4C). Claudins are tight junction membrane proteins that are expressed in epithelia and endothelia, and form paracellular barriers and pores that determine tight junction permeability (Gunzel & Yu, 2013). Earlier studies indicate that expression variation in claudins is important in permeability changes associated with salinity acclimation and possibly the formation of deeper tight junctions in the gills of freshwater fish (Bagherie-Lachidan, Wright, & Kelly, 2008; Kolosov, Bui, Chasiotis, & Kelly, 2013; Madsen & Tipsmark, 2008; Marshall et al., 2018; Tipsmark, Baltzegar, Ozden, Grubb, & Borski, 2008; Tipsmark et al., 2016). Our results suggest that both expression changes and genetic variation in *CLDND1* might play a key role in the repeated marine-freshwater divergence in nine-spined sticklebacks.

In conclusion, we used a population transcriptomic approach to uncover variation in both genetic and gene expression levels that is potentially associated with marine-freshwater divergence in nine-spined sticklebacks. Although a number of genes were identified to harbor SNPs associated with ecotypic differentiation in nine-spined sticklebacks, very few of these were shared with its close relative, the three-spined stickleback. Likewise, a number of genes were found to be differentially expressed between marine and freshwater nine-spined sticklebacks, several of which (12 of 861) are identified as candidates associated with marine-freshwater divergence in three-spined sticklebacks. However, few (e.g. *CLDND1*) seem to have been involved in repeated marine-freshwater divergence in nine-spined sticklebacks. Taken together, the results of this study

suggest that repeated marine-freshwater divergence in nine-spined sticklebacks is seldom underlain by similar genetic changes. The likely cause for this is the small effective population sizes of the populations studied here, as well as the likely polygenic nature of marine-freshwater divergence.

Acknowledgments

We thank Alexandre Budria, Per Byström, Chris Eberlein, and Ismo Rautiainen for help in fish collection, Kirsi Kähkönen for help in RNA extraction, Xinxin Li for help in data analyses, and Jacquelin DeFaveri for language checking. Logistic support from the Oulanka Biological Station (University of Oulu) is gratefully acknowledged. We are grateful for the computing resource support from CSC – the Finnish IT Center for Science Ltd administered by the Ministry of Education and Culture, Finland. B.G. thanks support from the National Natural Science Foundation of China (31672273) and CAS Pioneer Hundred Talents Program. This work is supported by the Academy of Finland (grant numbers, 129662, 134728 and 218343 to J.M.), and a grant from Helsinki Institute of Life Science (HiLIFE to J.M.).

Reference

Bagherie-Lachidan, M., Wright, S. I., & Kelly, S. P. (2008). Claudin-3 tight junction proteins in Tetraodon nigroviridis: cloning, tissue-specific expression, and a role in hydromineral balance. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology, 294*(5), R1638-R1647.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics, 7*(10), 781-791.

Berner, D., & Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics, 31*(9), 491-499.

Blouw, D. M., & Boyd, G. J. (1992). Inheritance of reduction, loss, and asymmetry of the pelvis in *Pungitius pungitius* (ninespine stickleback). *Heredity, 68*, 33-42.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., . . . Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature, 478*(7369), 343-348.

Bruneaux, M., Johnston, S. E., Herczeg, G., Merilä, J., Primmer, C. R., & Vasemagi, A. (2013). Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Molecular Ecology, 22*(3),

565-582.

Carroll, S. B. (2005). Evolution at two levels: On genes and form. *PLoS Biology, 3*(7), 1159-1166.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., . . . Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx*1 enhancer. *Science, 327*(5963), 302-305.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

DeFaveri, J., Shikano, T., Ab Ghani, N. I., & Merilä, J. (2012). Contrasting population structures in two sympatric fishes in the Baltic Sea basin. *Marine Biology, 159*(8), 1659-1672.

Earl, D. A., & Vonholdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources, 4*(2), 359-361.

Fang, B, Merilä, J., Matschiner, M., Momigliano, P. (2020). Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Molecular Phylogenetics and Evolution* 142, 106646

Fang, B., Kemppainen, P., Momigliano, P., & Merilä, J. (2019). Oceans apart: Heterogeneous patterns of parallel evolution in sticklebacks. *bioRxiv,* 826412.

Ferchaud, A. L., Pedersen, S. H., Bekkevold, D., Jian, J. B., Niu, Y. C., & Hansen, M. M. (2014). A low-density SNP array for analyzing differential selection in freshwater and marine populations of threespine stickleback (*Gasterosteus aculeatus*). *BMC Genomics, 15*, 863.

Foulkes, A. S. (2009). *Applied statistical genetics with R for population-based association studies*. Heidelberg: Springer.

Fu, J., Frazee, A. C., Collado-Torres, L., Jaffe, A. E., & Leek, J. T. (2019). ballgown: Flexible, isoform-level differential expression analysis. R package version 2.16.0.

Fu, Y., Wu, P. H., Beane, T., Zamore, P. D., & Weng, Z. P. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics, 19*, 531.

Gallant, J. R., Traeger, L. L., Volkening, J. D., Moffett, H., Chen, P. H., Novina, C. D., . . . Sussman, M. R. (2014). Genomic basis for the convergent evolution of electric organs. *Science, 344*(6191), 1522-1525.

Gibbons, T. C., Metzger, D. C. H., Healy, T. M., & Schulte, P. M. (2017). Gene expression plasticity in response to salinity acclimation in threespine stickleback ecotypes from

different salinity habitats. *Molecular Ecology, 26*(10), 2711-2725.

Gonda, A., Herczeg, G., & Merilä, J. (2009). Adaptive brain size differentiation in ninespine sticklebacks. *Journal of Evolutionary Biology*, *22*, 1721–1726.

Gonda, A., Herczeg, G., & Merilä, J. (2011). Population variation in brain size of nine-spined sticklebacks (*Pungitius pungitius*): local adaptation or environmentally induced variation? *BMC Evolutionary Biology*, *11*, 75.

Gunzel, D., & Yu, A. S. L. (2013). Claudins and the modulation of tight junction permeability. *Physiological Reviews, 93*(2), 525-569.

Guo, B., Chain, F. J. J., Bornberg-Bauer, E., Leder, E. H., & Merilä, J. (2013). Genomic divergence between nine- and three-spined sticklebacks. *BMC Genomics, 14*, 756.

Guo, B., Fang, B., Shikano, T., Momigliano, P., Wang, C., Kravchenko, A., & Merilä, J. (2019). A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus *Pungitius*. *Molecular Ecology*, 28(17), 4046-4064.

Hao, Y., Xiong, Y., Cheng, Y. L., Song, G., Jia, C. X., Qu, Y. H., & Lei, F. M. (2019). Comparative transcriptomics of 3 high-altitude passerine birds and their low-altitude relatives. *Proceedings of the National Academy of Sciences of the United States of America, 116*(24), 11851-11856.

Hart, J. C., Ellis, N. A., Eisen, M. B., & Miller, C. T. (2018). Convergent evolution of gene expression in two high-toothed stickleback populations. *PLoS Genetics, 14*(6), e1007443.

Healy, T. M., Brennan, R. S., Whitehead, A., & Schulte, P. M. (2018). Tolerance traits related to climate change resilience are independent and polygenic. *Global Change Biology, 24*(11), 5348-5360.

Herczeg, G., Turtiainen, M., & Merilä, J. (2010). Morphological divergence of North-European nine-spined sticklebacks (*Pungitius pungitius*): signatures of parallel evolution. *Biological Journal of the Linnean Society, 101*(2), 403-416.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics, 6*(2), e1000862.

Hu, Y. B., Wu, Q., Ma, S., Ma, T. X., Shan, L., Wang, X., . . . Wei, F. W. (2017). Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proceedings of the National Academy of Sciences of the United States of America, 114*(5), 1081-1086.

Ishikawa, A., Kabeya, N., Ikeya, K., Kakioka, R., Cech, J. N., Osada, N., . . . Kitano, J. (2019). A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science, 364*(6443), 886-889.

Ishikawa, A., Kusakabe, M., Yoshida, K., Ravinet, M., Makino, T., Toyoda, A., . . . Kitano, J. (2017). Different contributions of local- and distant-regulatory changes to transcriptome divergence between stickleback ecotypes. *Evolution, 71*(3), 565-581.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics, 23*(14), 1801-1806.

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., . . . Kingsley, D. M. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology, 22*(1), 83-90.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature, 484*(7392), 55-61.

Kemppainen, P., Li, Z., Rastas, P., Löytynoja, A., Fang, B., Guo, B., Shikano, T., Yang, J., & Merilä, J. (2020). Convergent genetic architecture underlies parallel pelvic reduction in a stickleback species with high population structuring and isolation by distance. bioRxiv:2020.2001.2017.908970.

King, M., & Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science, 188*(4184), 107-116.

Klepaker, T., Ostbye, K., & Bell, M. A. (2013). Regressive evolution of the pelvic complex in stickleback fishes: a study of convergent evolution. *Evolutionary Ecology Research, 15*(4), 413-435.

Kolosov, D., Bui, P., Chasiotis, H., & Kelly, S. P. (2013). Claudins in teleost fishes. *Tissue Barriers, 1*(3), e25391.

Kusakabe, M., Ishikawa, A., Ravinet, M., Yoshida, K., Makino, T., Toyoda, A., . . . Kitano, J. (2017). Genetic basis for variation in salinity tolerance between stickleback ecotypes. *Molecular Ecology, 26*(1), 304-319.

Laine, V. N., Shikano, T., Herczeg, G., Vilkki, J., & Merilä, J. (2013). Quantitative trait loci for growth and body size in the nine-spined stickleback *Pungitius pungitius* L. *Molecular Ecology, 22*(23), 5861-5876.

Leder, E. H., McCairns, R. J., Leinonen, T., Cano, J. M., Viitaniemi, H. M., Nikinmaa, M., . . . Merilä, J. (2015). The evolution and adaptive potential of transcriptional variation in sticklebacks--signatures of selection and widespread heritability. *Molecular Biology and Evolution, 32*(3), 674-689.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Proc, G. P. D. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li, Z., Loytynoja, A., Fraimout, A., & Merila, J. (2019). Effects of marker type and filtering criteria on Q_{ST}-F_{ST} comparisons. *Royal Society Open Science, 6*(11), 190666.

Madsen, S. S., & Tipsmark, C. K. (2008). Changes in claudin isoform expression in the gill during salinity shifts and smoltification of Atlantic salmon. *FASEB Journal, 22,* 1200.

Marshall, W. S., Breves, J. P., Doohan, E. M., Tipsmark, C. K., Kelly, S. P., Robertson, G. N., & Schulte, P. M. (2018). claudin-10 isoform expression and cation selectivity change with salinity in salt-secreting epithelia of *Fundulus heteroclitus*. *Journal of Experimental Biology, 221*(1).

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462.

Merilä, J. (2013). Nine-spined stickleback (*Pungitius pungitius*): an emerging model for evolutionary biology research. *Annals of the New York Academy of Sciences, 1289*, 18-35.

Merilä, J. (2014). Lakes and ponds as model systems to study parallel evolution. *Journal of Limnology*, 73(s1): 33-45.

Ogura, A., Ikeo, K., & Gojobori, T. (2004). Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human. *Genome Research, 14*(8), 1555-1561.

Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A., & Oakley, T. H. (2014). Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proceedings of the National Academy of Sciences of the United States of America, 111*(44), E4736-E4742.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, *6*, 25533.

Peichel, C. L., Nereng, K. S., Ohgi, K. A., Cole, B. L. E., Colosimo, P. F., Buerkle, C. A., . . . Kingsley, D. M. (2001). The genetic architecture of divergence between threespine stickleback species. *Nature, 414*(6866), 901-905.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols, 11*(9), 1650-1667.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Pritchard, V. L., Viitaniemi, H. M., McCairns, R. J. S., Merilä, J., Nikinmaa, M., Primmer, C. R., & Leder, E. H. (2017). Regulatory architecture of gene expression variation in the threespine stickleback *Gasterosteus aculeatus*. *G3-Genes Genomes Genetics, 7*(1), 165-178.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559-575.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842.

Raeymaekers, J. A. M., Chaturvedi, A., Hablutzel, P. I., Verdonck, I., Hellemans, B., Maes, G. E., . . . Volckaert, F. A. M. (2017). Adaptive and non-adaptive divergence in a common landscape. *Nature Communications, 8*, 267.

Rastas, P., Calboli, F. C., Guo, B., Shikano, T., & Merilä, J. (2016). Construction of ultradense linkage maps with Lep-MAP2: stickleback F_2 recombinant crosses as an example. *Genome Biology and Evolution, 8*(1), 78-93.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes, 4*(1), 137-138.

Rosenblum, E. B., Parent, C. E., & Brandt, E. E. (2014). The molecular vasis of phenotypic convergence. *Annual Review of Ecology, Evolution, and Systematics, 45*, 203-226.

Shapiro, M. D., Bell, M. A., & Kingsley, D. M. (2006). Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America, 103*(37), 13753-13758.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., . . . Kingsley, D. M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature, 428*(6984), 717-723.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., . . . Kingsley, D. M. (2006). Genetic and developmental basis of evolutionary pelvic reduction

in threespine sticklebacks. *Nature, 439*(7079), 1014-1014.

Shapiro, M. D., Summers, B. R., Balabhadra, S., Aldenhoven, J. T., Miller, A. L., Cunningham, C. B., . . . Kingsley, D. M. (2009). The genetic architecture of skeletal convergence and sex determination in ninespine sticklebacks. *Current Biology, 19*(13), 1156-1156.

Shikano, T., Laine, V. N., Herczeg, G., Vilkki, J., & Merilä, J. (2013). Genetic architecture of parallel pelvic reduction in ninespine sticklebacks. *G3-Genes Genomes Genetics, 3*(10), 1833-1842.

Shikano, T., Shimada, Y., Herczeg, G., & Merilä, J. (2010). History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations. *Molecular Ecology, 19*(6), 1147-1161.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics, 20*(11), 631-656.

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics, 14*(11), 751-764.

Tipsmark, C. K., Baltzegar, D. A., Ozden, O., Grubb, B. J., & Borski, R. J. (2008). Salinity regulates claudin mRNA and protein expression in the teleost gill. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology, 294*(3), R1004-R1014.

Tipsmark, C. K., Breves, J. P., Rabeneck, D. B., Trubitt, R. T., Lerner, D. T., & Grau, E. G. (2016). Regulation of gill claudin paralogs by salinity, cortisol and prolactin in Mozambique tilapia (*Oreochromis mossambicus*). *Comparative Biochemistry and Physiology A-Molecular & Integrative Physiology, 199*, 78-86.

Varadharajan, S., P. Rastas, A. Loytynoja, M. Matschiner, F. C. F. Calboli, B. Guo, A. J. Nederbragt, K. S. Jakobsen, and J. Merilä. 2019. A high-quality assembly of the nine-spined stickleback (*Pungitius pungitius*) genome. *Genome Biology and Evolution* 11(11):3291-3308.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols, 11*(1), 1-9.

Wang, G., Yang, E., Smith, K. J., Zeng, Y., Ji, G. L., Connon, R., . . . Cai, J. J. (2014). Gene expression responses of threespine stickleback to salinity: implications for salt-sensitive hypertension. *Frontiers in Genetics, 5*, 312.

Wootton, R. J. (1976). *The biology of the sticklebacks*. New York: Academic.

Xie, K. T., Wang, G. L., Thompson, A. C., Wucherpfennig, J. I., Reimchen, T. E., MacColl, A. D.

C., . . . Kingsley, D. M. (2019). DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science, 363*(6422), 81-84.

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols, 10*(10), 1556-1566.

Yu, G. C., Wang, L. G., Han, Y. Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics-a Journal of Integrative Biology, 16*(5), 284-287.

Zou H and Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67: 301–320.

Data Accessibility

RNA-seq sequences underlying this study have been deposited in NCBI's Sequence Read Archive and accession numbers are listed in Table S1.

Author Contributions

BG and JM conceived the project. Yingnan Wang, YZ, ZL, Yu Wang, and BG analyzed the data. BG, Yingnan Wang, ZL, and JM wrote the paper. All authors read and approved the final manuscript.

Figure legends

Fig. 1 (A) Map of the Fennoscandia, showing the locations of the nine-spined stickleback populations used in this study. 'SWE-ABB' = Abbortjärn pond, Sweden; 'SWE-BYN' = Bynästjärnen pond, Sweden; 'SWE-BOL' = Baltic Sea at Bölesviken, Sweden; 'FIN-PYO' = Pyöreälampi pond, Finland; 'FIN-RYT' = Rytilampi pond, Finland; 'FIN-HEL' = Baltic Sea at Helsinki, Finland. (B) Principal Component Analysis of autosomal SNPs. Principal components (PCs) 1, 2, and 3 are shown. (C) Unrooted neighbor-joining tree based on identity by state distance matrix of autosomal SNPs. The six populations are divided into three lineages, marine lineage with both Finnish and Swedish marine populations, Finnish pond lineage, and Swedish pond lineage, with a bootstrap value of 100. Marine populations are marked with blue arc line and freshwater populations with purple arc line. (D) Genetic clustering with autosomal SNPs. The number of populations (K) was predefined from 2 to 6, with the best fit scenario of $K = 3$.

Fig. 2 Genome-wide distribution of genetic variation and differentially expressed genes between marine and freshwater nine-spined stickleback populations. Linkage groups are labeled in black Arabic numerals and represented as grey blocks in the circle. All identified bi-allelic SNPs (green), SNPs associated with marine-freshwater divergence (dark green), genes with SNPs associated with marine-freshwater divergence (blue), and differentially expressed genes (DEGs) in the brain (purple) and in the liver (red) are plotted as occurrence density functions in genomic position with a non-overlapping 2Mb sliding window.

Fig. 3 Significantly enriched GO terms of differentially expressed genes/transcripts (DEGs/DETs) identified with the pooled approach, in which gene expression profiles are compared quantitatively between marine and freshwater ecotypes by pooling nine-spined sticklebacks from the same ecotype into a group. (A) DEGs in the brain; (B) DEGs in the liver; (C) DETs in the brain; (D) DETs in the liver.

Fig. 4 *CLDND1* – a candidate gene potentially associated with repeated marine-freshwater divergence in nine-spined sticklebacks. (A) Position, alleles, mutation type, allele frequency of each SNP in *CLDND1* gene in each of the six nine-spined populations; (B) location of each SNP on the secondary structure of CLDND1 protein, (C) expression quantity of *CLDND1* gene in each individual of the six nine-spined populations.

Tables

Table 1 Differentially expressed genes/transcripts (DEGs/DETs) potentially associated with repeated marine-freshwater divergence in nine-spined sticklebacks

Table 2 Genes associated with repeated marine-freshwater divergence in nine-spined sticklebacks with high confidence.

Supporting information

Fig. S1 Sequencing reproducibility of each of the 48 transcriptomes. Each transcriptome was sequenced in two different sequence lanes (*viz.* lane 1 and lane 2 for brain, and lane 7 and lane 8 for liver).

Fig. S2 Significantly enriched GO terms of genes harboring SNPs associated with marine-freshwater divergence based on GWAS.

Fig. S3 Principal component analysis of gene expression in the brain (A) and liver (B). The first and second PCs are plotted as in the X- and Y- axes, respectively.

Fig. S4 Heat map of cosine similarity between each pair of transcriptomes in the brain (left panel) and liver (right panel). A total of 9,752 genes found in all of the 24 brain transcriptomes with TPM ≥ 1 were used in the pairwise cosine similarity in the brain transcriptome comparison, and 5,131 genes were used in the pairwise cosine similarity in the brain transcriptome comparison.

Fig. S5 Venn diagram illustrating numbers of differentially expressed genes/transcripts (DEGs/DETs) identified with a pooled approach.

Fig. S6 Interactions among identified DEGs associated with marine-freshwater divergence in the brain with a pooled approach. DEGs with known names are shown.

Fig. S7 Venn diagram illustrating numbers of differentially expressed genes/transcripts (DEGs/DETs) identified with the integrated approach.

Fig. S8 Significantly enriched GO terms of differentially expressed genes/transcripts (DEGs/DETs) identified with an integrated approach in nine-spined sticklebacks. (A) DEGs in the brain; (B) DEGs in the liver; (C) DETs in the brain; (D) DETs in the liver.

Table S1 Information on samples used in this study. Read numbers refer to sequenced short reads for each transcriptome, and accession number is the unique identifier of each transcriptome in GenBank.

Table S2 Distribution of identified SNPs across the nine-spined stickleback genome

Table S3 Genes with candidate SNPs associated with divergence between marine and freshwater sticklebacks

Table S4 Differentially expressed genes/transcripts (DEGs/DETs) identified with a pooled approach

Table S5 Clusters include no less than 10 genes showing strong interactions among each other. Genes identified as DEGs with a pooled approach are highlighted in yellow

Table S6 Differentially expressed genes/transcripts (DEGs/DETs) identified with an integrated approach

Table 1 Differentially expressed genes/transcripts (DEGs/DETs) potentially associated with repeated marine-freshwater divergence in nine-spined sticklebacks

Frequency	Brain		Liver	
	DEGs	DETs	DEGs	DETs
1	525	636	668	222
2	86(**3**)	119(**1**)	11(**1**)	5(**0**)
3	22(**7**)	22(**4**)	0	1(**0**)
4	3(**2**)	19(**2**)	0	0
5	1(**1**)	2(**2**)	0	0
6	1(**1**)	2(**2**)	0	0
7	0	0	0	0
8	0	1(**1**)	0	0
Total	638(**14**)	801(**12**)	679(**1**)	228(**0**)

Note: Numbers in brackets are DEGs/DETs that appear in ≥ two pairwise marine and freshwater comparisons that are not between one marine and two freshwater populations or vice versa.

Table 2 Genes associated with repeated marine-freshwater divergence in nine-spined sticklebacks with high confidence.

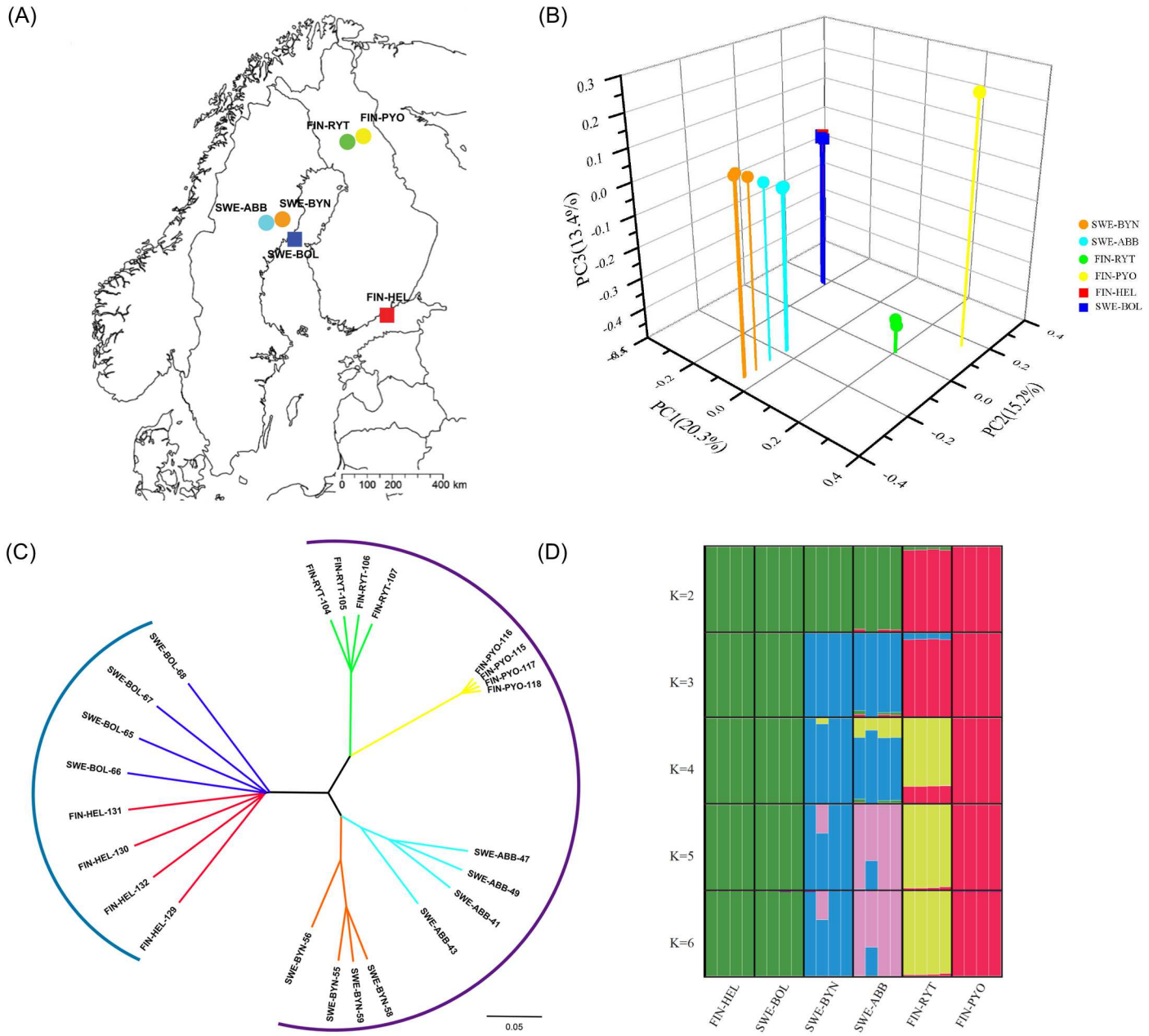
Gene ID	Gene name	Chromosome	Gene start	Gene end	strand	Gene description	GO term	Three-spined stickleback	Reference
DEGs/DETs identified with pooled approach and potentially associated with repeated adaptation from marine to freshwater in brain according to integrated approach									
MSTRG.22266	Cacybp	Chr3	11963497	11966768	-	Calcyclin-binding protein			
MSTRG.26830	cfl2	Chr6	11815136	11817458	+	Cofilin-2	GO:0003779 GO:0005622 GO:0015629 GO:0030042	√	Gibbons et al. 2017
MSTRG.30600	Pprc1	Chr9	9278954	9285388	+	Peroxisome proliferator-activated receptor gamma coactivator-related protein 1	GO:0000166 GO:0003676	√	Wang et al. 2014; Gibbons et al. 2017

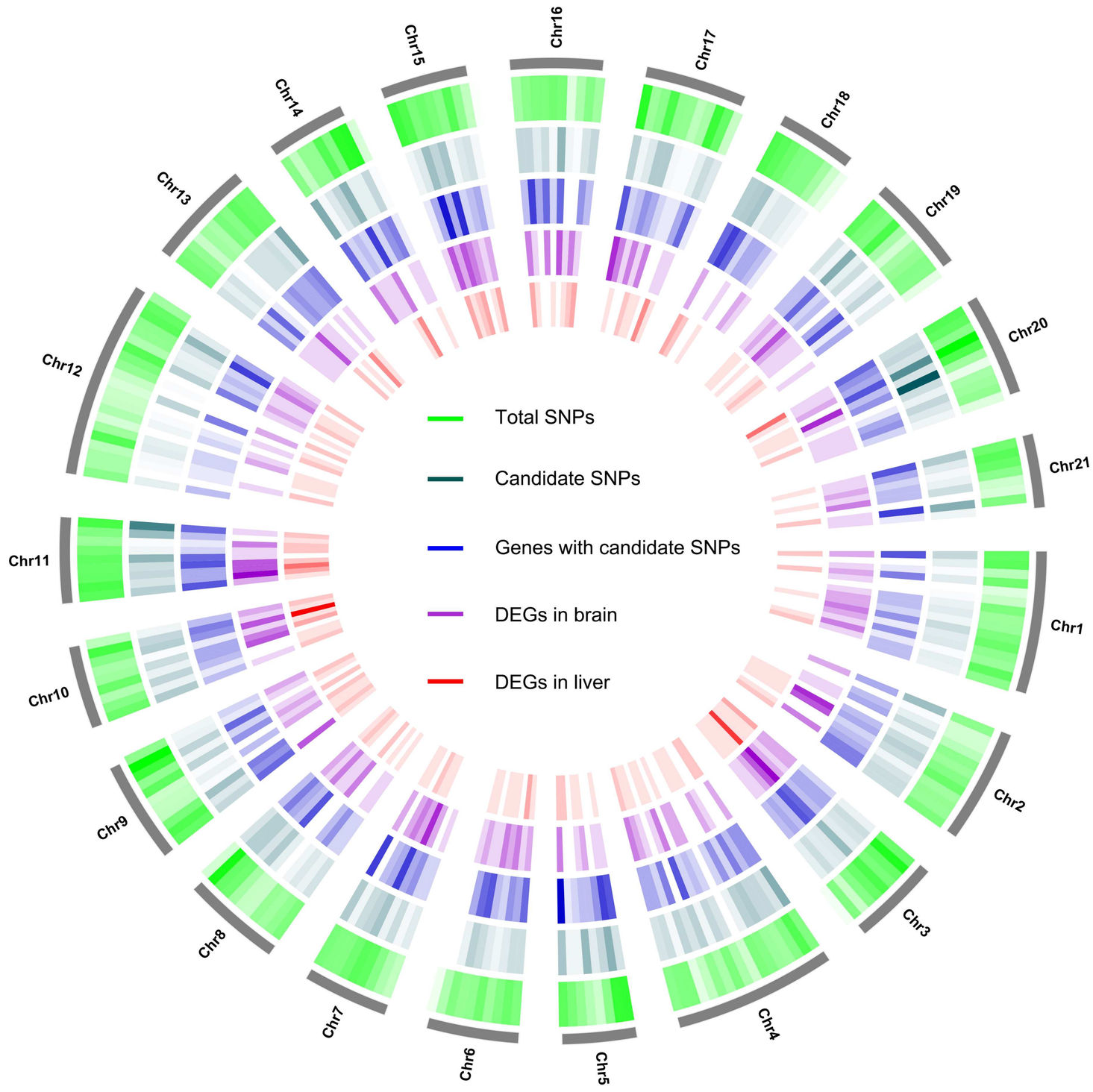
Ppun_00029742-RA	Acadm	Chr3	7072375	7078454	-	Medium-chain specific acyl-CoA dehydrogenase, mitochondrial	GO:0003995 GO:0008152 GO:0016627 GO:0050660 GO:0055114		
MSTRG.25445.5	OAT	Chr5	6374753	6379835	-	Ornithine aminotransferase, mitochondrial	GO:0003824 GO:0008483 GO:0030170	√	Gibbons et al. 2017
MSTRG.27871	sod1	Chr7	10027033	10029462	+	Superoxide dismutase [Cu-Zn]	GO:0004784 GO:0006801 GO:0046872 GO:0055114		
DEGs/DETs identified with pooled approach and harboring SNPs associated with marine-freshwater divergence according to GWAS									
MSTRG.11240	Trnp1	Chr15	8861127	8861744	+	TMF-regulated nuclear protein 1	GO:0005515		
MSTRG.1154	GEMIN7	Chr1	19557626	19560111	-	Gem-associated protein 7			
MSTRG.11579	TTC7B	Chr15	14444492	14462593	+	Tetratricopeptide repeat protein 7B	GO:0005515	√	Gibbons et al. 2017
MSTRG.12869	sestd1(2)	Chr16	16716857	16734065	+	SEC14 domain and spectrin repeat-containing protein 1			
MSTRG.18345	MADD	Chr2	21429917	21464726	-	MAP kinase-activating death domain protein			
MSTRG.24190	Ube2k	Chr4	23555896	23559686	+	Ubiquitin-conjugating enzyme E2 K	GO:0005515	√	Wang et al. 2014

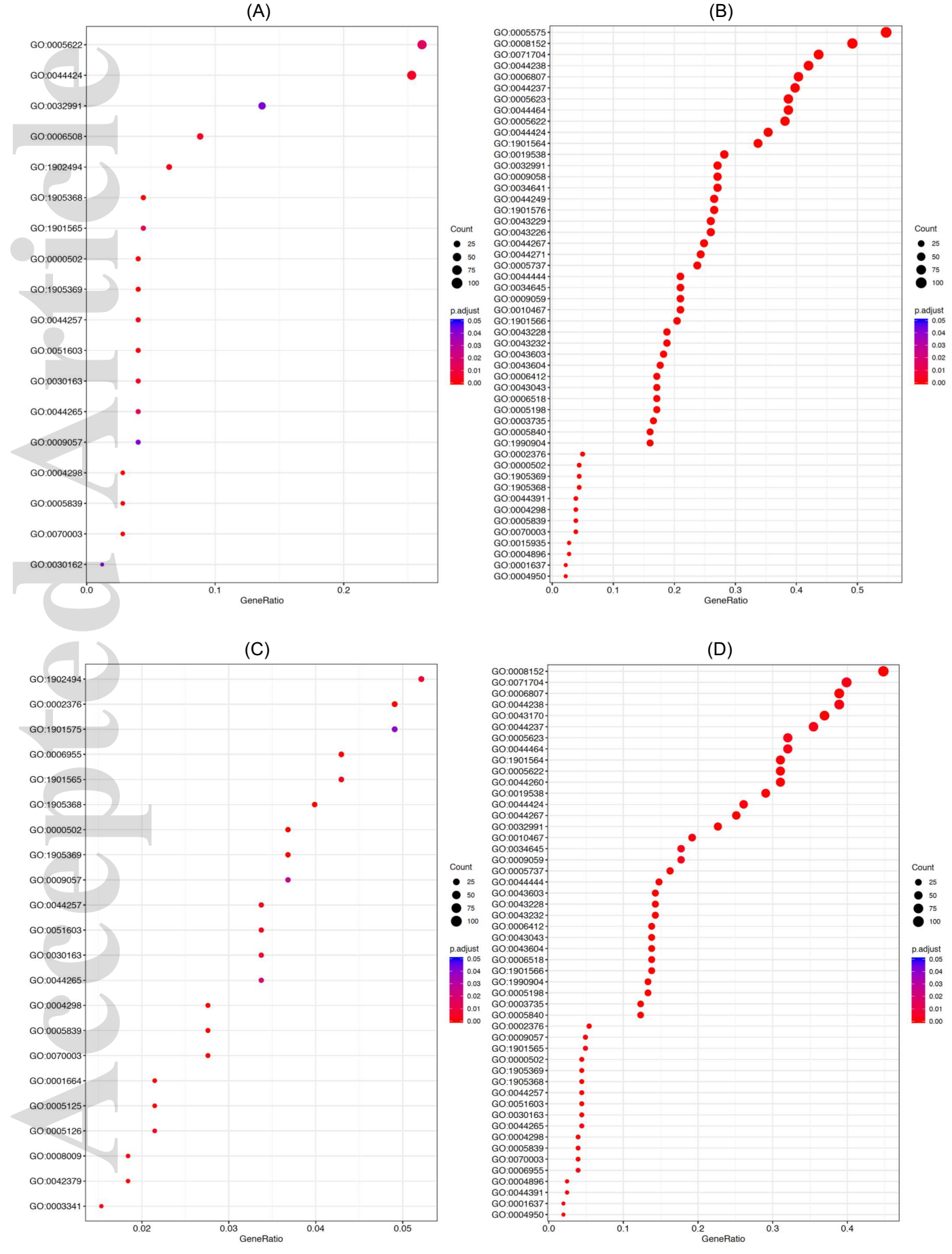
MSTRG.24775	sept8a	Chr4	32722297	32729267	+	Septin-8-A	GO:0005525		
MSTRG.25613	Protein of unknown function	Chr5	9404766	9420059	-				
MSTRG.26227	Plekha3	Chr6	2150395	2154781	-	Pleckstrin homology domain-containing family A member 3			
MSTRG.4412	SCAF1	Chr11	13791168	13799043	-	Splicing factor, arginine/serine-rich 19			
MSTRG.6811	PELP1	Chr12	29646548	29658544	+	Proline-, glutamic acid- and leucine-rich protein 1	GO:0005488	√	Gibbons et al. 2017
MSTRG.7080	CLDND1	Chr12	35236933	35239625	+	Claudin domain-containing protein 1	GO:0016021		
MSTRG.8008	PPP3CC	Chr13	5131675	5150357	+	Serine/threonine-protein phosphatase 2B catalytic subunit gamma isoform	GO:0016787		
MSTRG.9355	Protein of unknown function	Chr14	1081087	1084322	-				
MSTRG.961	Diablo	Chr1	16578339	16580991	-	Diablo homolog, mitochondrial	GO:0005739 GO:0006915 GO:0006919	√	Wang et al. 2014
MSTRG.21985	Ripk2	Chr3	8575796	8582156	-	Receptor-interacting serine/threonine-protein kinase 2	GO:0004672 GO:0005524 GO:0006468	√	Gibbons et al. 2017

	GO:0042981

Note: Genes that have been identified to be associated with marine-freshwater divergence in the three-spined stickleback are marked with "√".



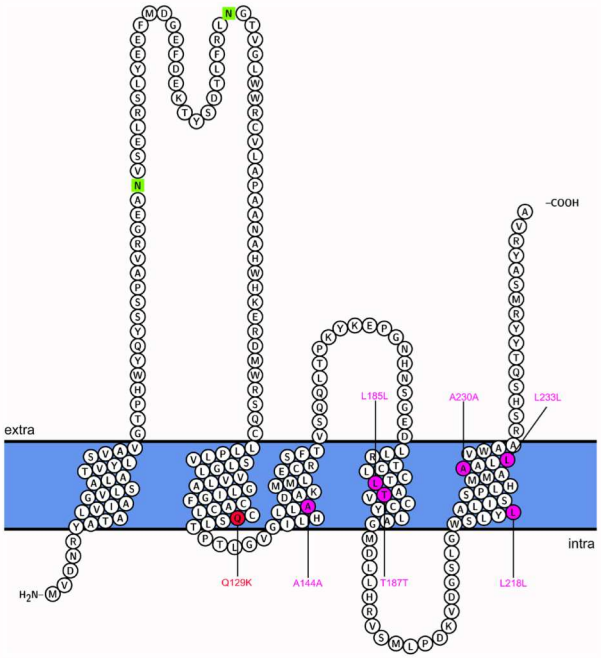




(A)

Order	Chromosome	Position	Marine	Freshwater	Mutation_Type	Protein_Change	FIN-HEL	SWE-BOL	FIN-PYO	FIN-RYT	SWE-ABB	SWE-BYN
1	Chr12	35238171	C	A	Nonsynonymous	Q129K						
2	Chr12	35238218	C	T	Synonymous	A144A						
3	Chr12	35238622	G	A	Synonymous	L185L						
4	Chr12	35238628	A	C	Synonymous	T187T						
5	Chr12	35238719	T	C	Synonymous	L218L						
6	Chr12	35238757	T	C	Synonymous	A230A						
7	Chr12	35238766	T	C	Synonymous	L233L						

(B)



(C)